Pitch Velocity vs Performance

Introduction

"These guys today are just throwers, not pitchers." "You have to change speeds to throw hitters off." "The slower your changeup is, the faster your fastball is." These are common clichés used in baseball to express a pitchers ability to throw pitches at different speeds is just as, if not more important than his ability to throw fast. In all fairness, there is a lot more to pitching than just velocity, such as how much a pitch breaks, or how well the pitcher can control his pitches. However, if you were going to try to predict a pitchers success using no more than a radar gun, should you simply judge how hard a pitcher throws, or should you consider the range of velocity a pitcher throws in?

Purpose

The purpose of this study is to see if a relationship exists between a pitchers success, average velocity, and range of velocity, and if a relationship exists does average velocity or range of velocity have a stronger relationship than the other.

Method Overview

To look at the relationship of velocity vs success, I will be using R programming to produce scatter plots, perform principal components analysis, and produce Generalized Linear Models. The predictor variables in this study will be average fastball velocity, range of fastball velocity, average breaking ball velocity, range of breaking ball velocity, average off speed pitch velocity, range of off speed pitch velocity, the difference between average fastball velocity and average breaking ball velocity, the difference between average fastball velocity and average off speed pitch velocity, and the difference between average off speed pitch velocity and average breaking ball velocity. The response variable to measure success in this study will be Earned Run Average (ERA).

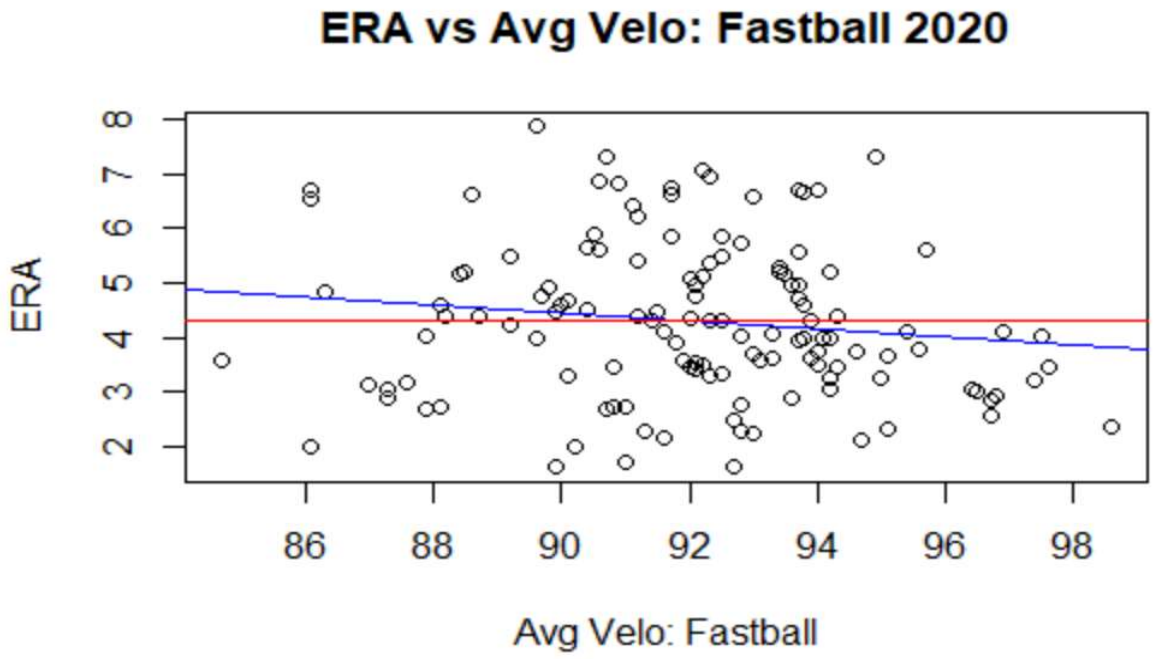
Data

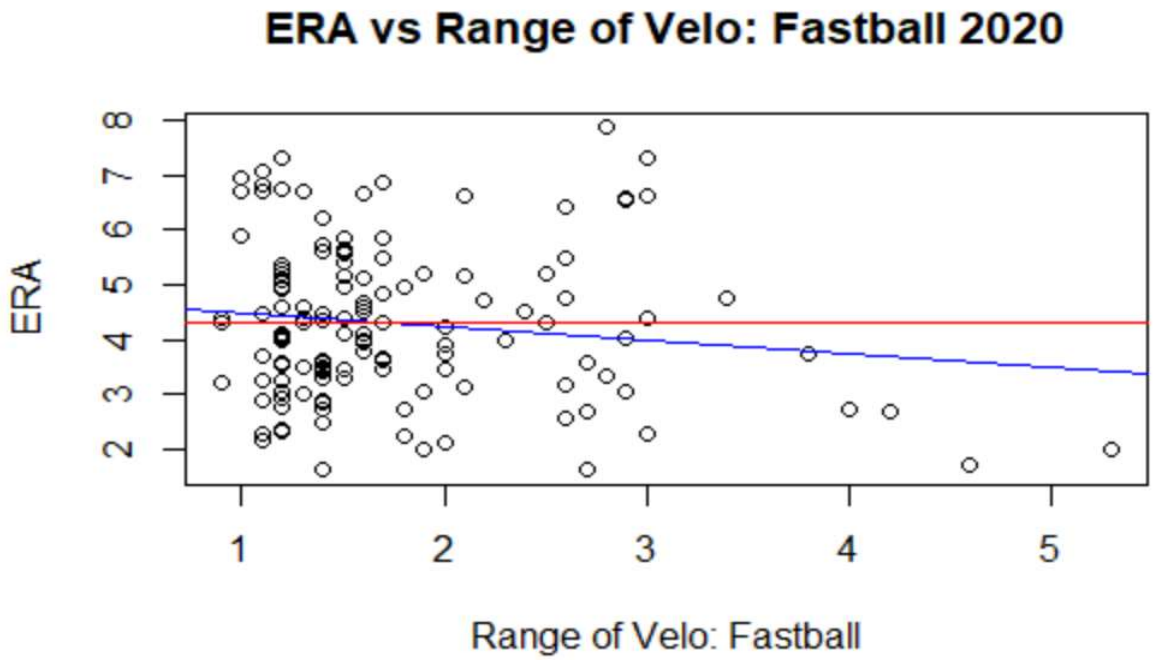
All data was gathered from the website Baseball Savant. The training data comes from the 2020 MLB season, and 2019 will be used as test data to ensure the results uphold overtime. It is worth noting, due to COVID-19 the 2020 season was shortened to 60 games. With that there is a smaller sample size of this data as opposed to a typical full 162 game season, which may produce more predictive uncertainty than normal.

It is important to note, the main intention of this study is to focus on starting pitchers who would throw multiple innings at a time. For that reason, the pitchers included in this study are all pitchers that faced a minimum of 150 batters in 2020. Making the selection of 150 batters faced, does include a handful of 'swing men' that may have bounced between the bullpen and starting rotation. However, seeing how most of these pitchers are still going to be throwing multiple innings at a time in a typical appearance, and the definition of pitching roles is ever changing in today's game, the role of these pitchers is similar enough to a starter that it is appropriate to include them. In addition, there were only 40 pitchers who threw enough innings to qualify as a starter, which would produce a small sample size. Expanding the definition to 150 batters faced gave a total of 133 pitchers, or observations in the training data. On the contrary, making the same selection for pitchers in 2019 would have included too many relievers, due to having a full season, which may have skewed the data. As a result the 2019 data set is all pitchers that qualified for the ERA title.

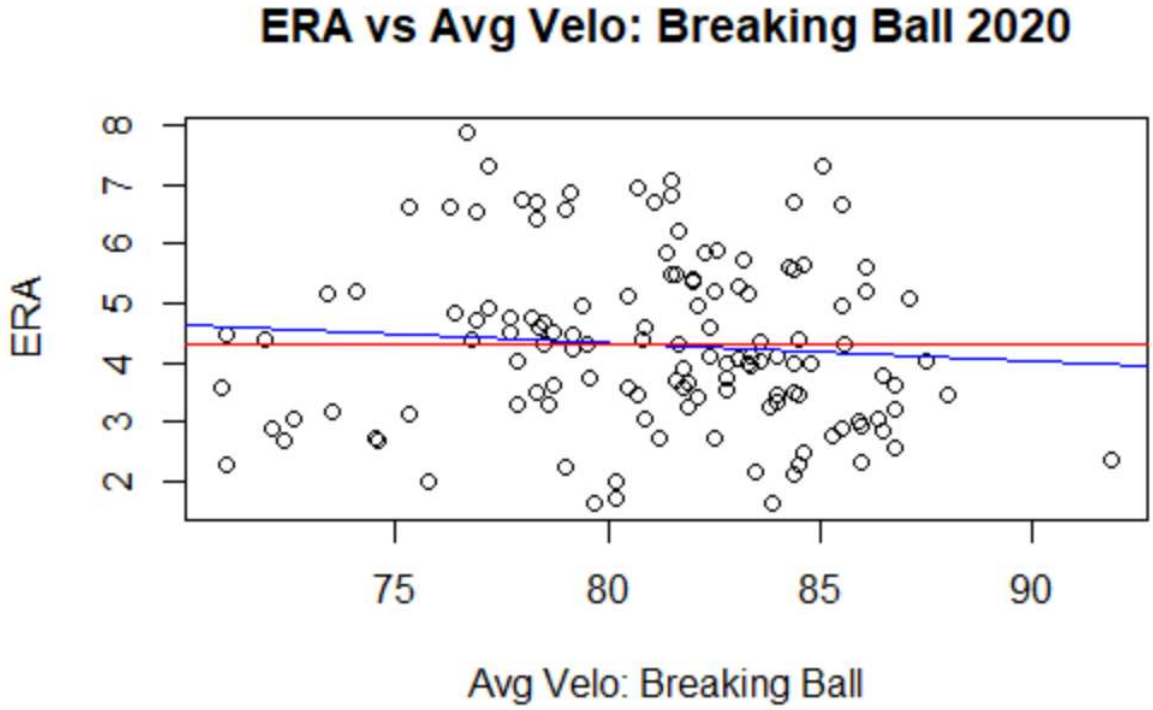
Another note about the data is pitch definition. Some pitchers have multiple pitches that may be defined as a fastball, off speed pitch, or breaking ball. Other pitchers may only have one of a given category. This may come into play by making it look like a pitcher was a wider range of velocity on a particular pitch relative to another pitcher, but in reality it is because that range comes from a difference between two separate pitchers as opposed to one. Considering one of the intentions of this study is to evaluate the significance of mixing pitches, there will be no adjustment for a pitcher that throws multiple pitches of the same type. In addition, not every pitcher has every type of pitch. In particular some pitchers only throw fastballs and breaking balls, but no off speed pitches. These pitchers will be removed for the purposed of this study. In 2020 Garret Richard and Dinelson Lamet were removed for this reason. In 2019 Robbie Ray and Joey Lucchesi were removed.

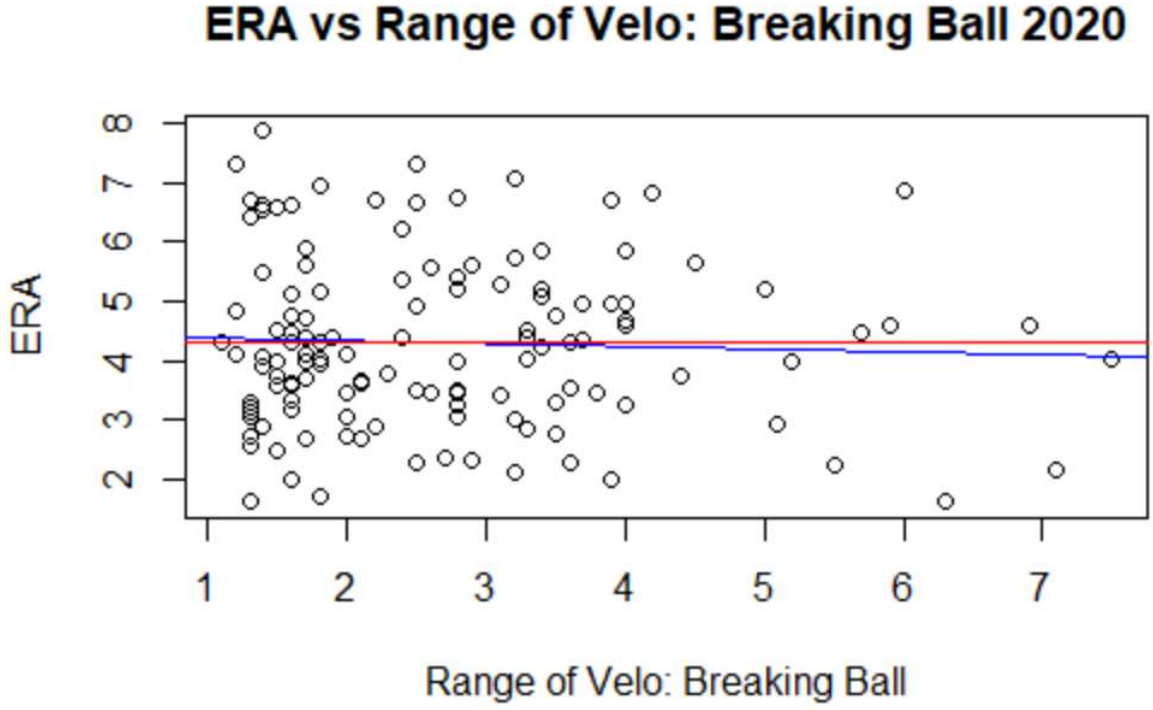
Scatter Plots

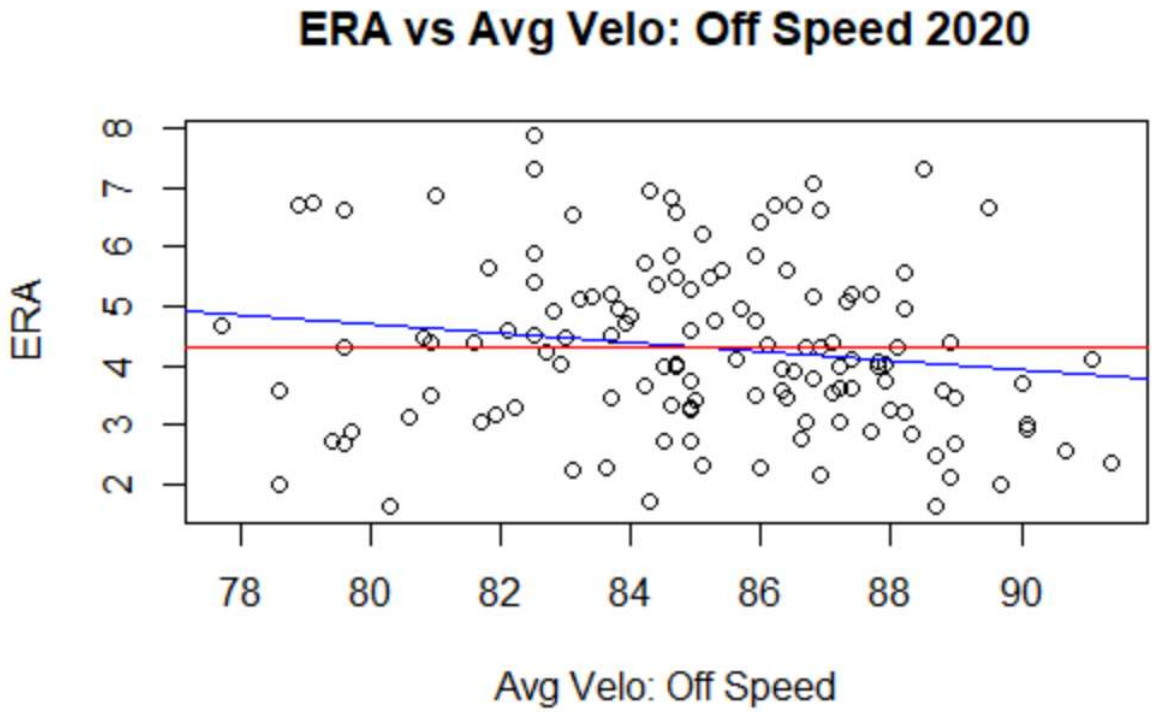


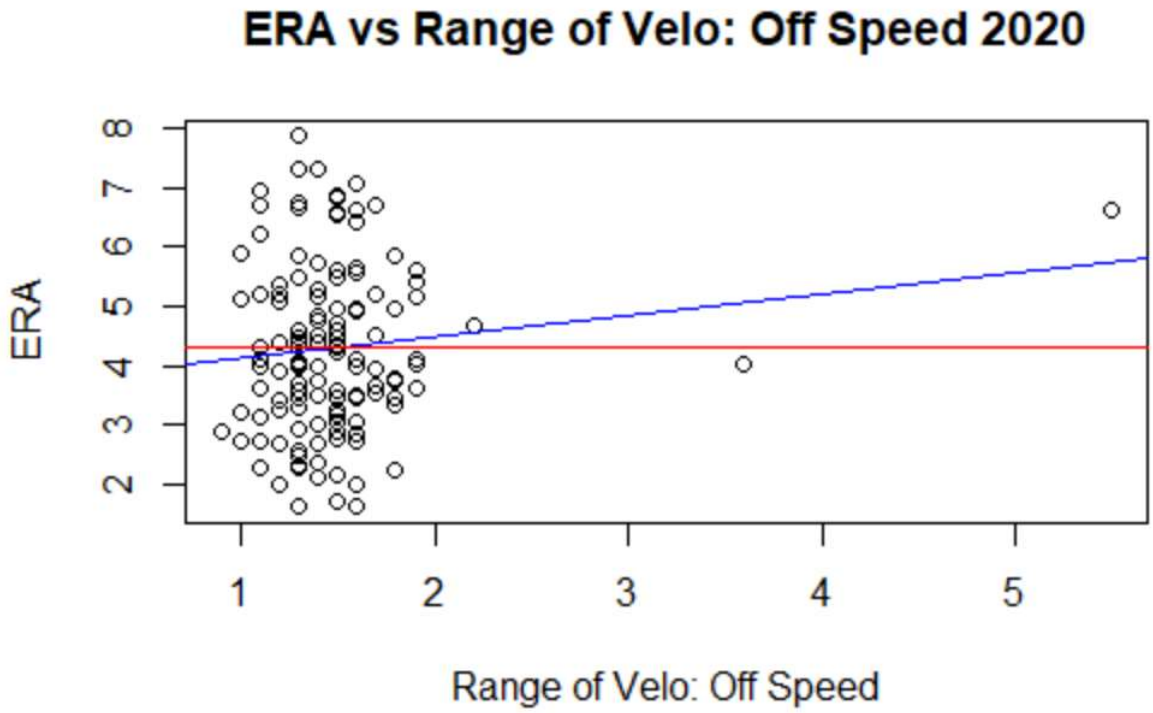


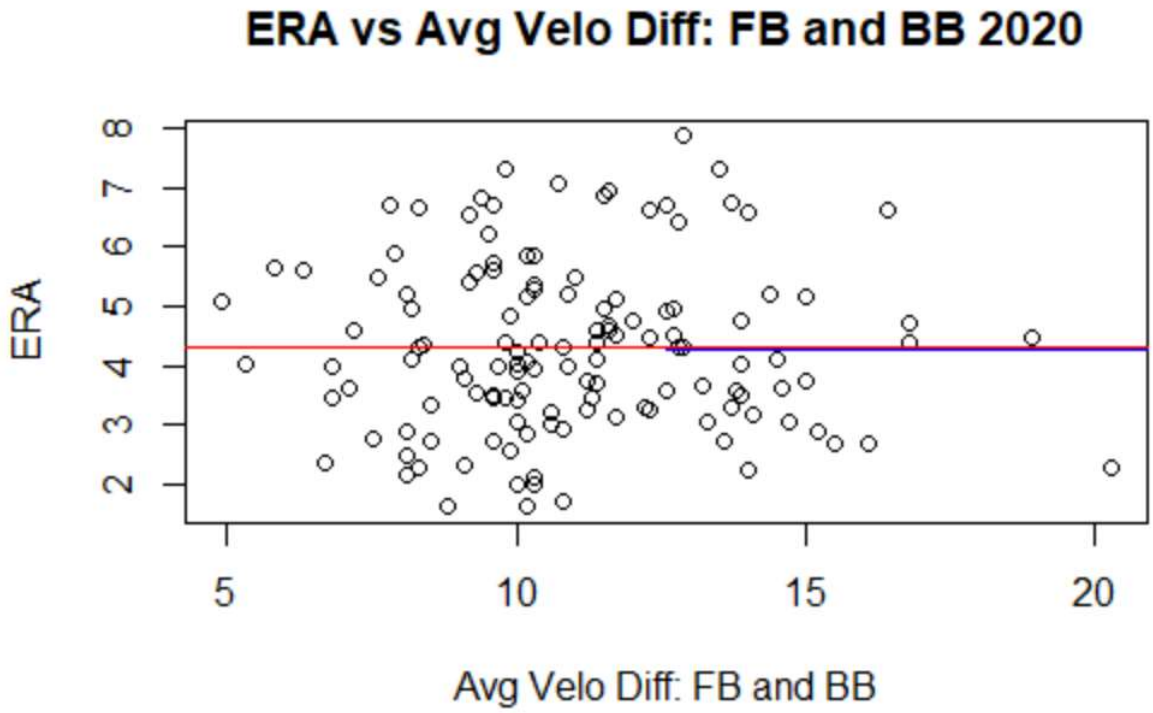
*Red Line represents the sample mean
**Blue Line represents the ordinary least squares best fit line

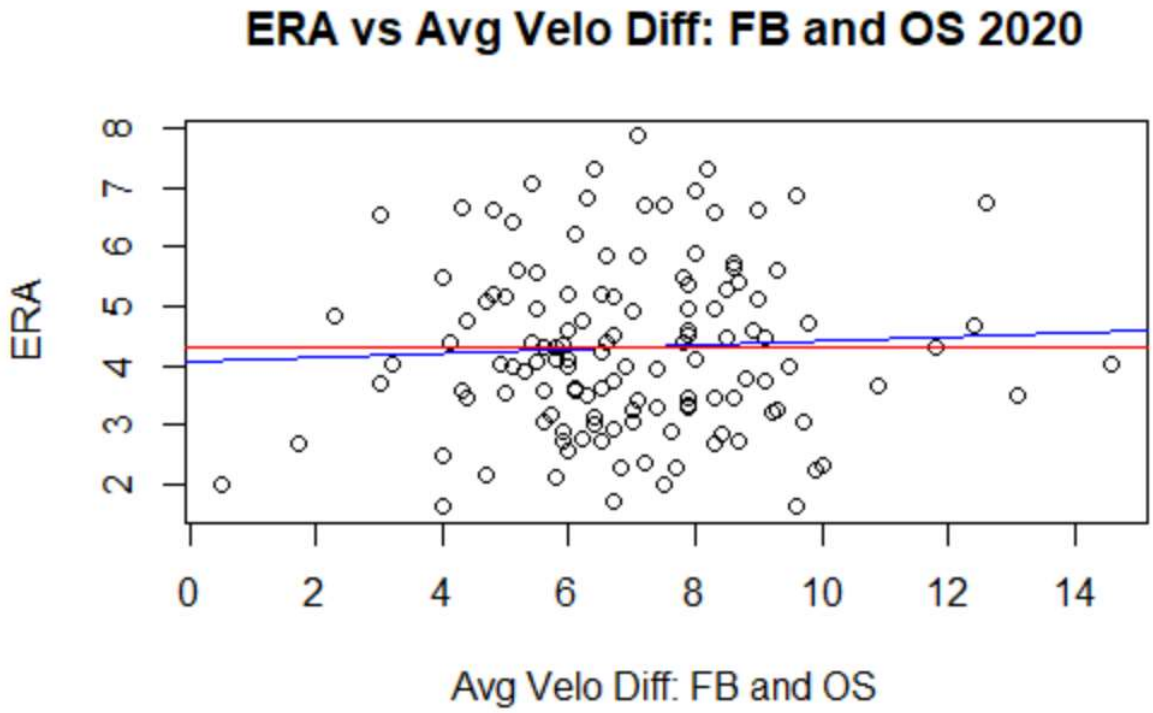

ERA vs Avg Velo: Breaking Ball 2020
ERA
2
3
4
5
6
7
8
75
80
85
90
Avg Velo: Breaking Ball

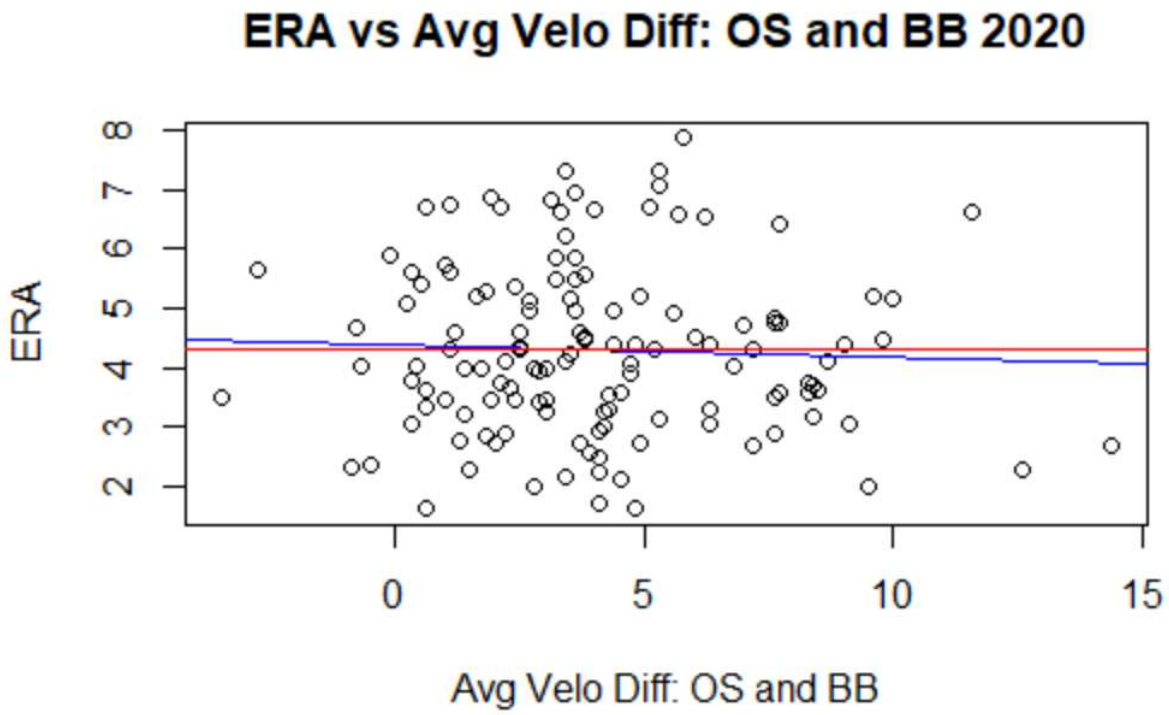

ERA vs Range of Velo: Breaking Ball 2020
ERA
2
3
4
5
6
7
8
1
2
3
4
5
6
7
Range of Velo: Breaking Ball


ERA vs Avg Velo: Off Speed 2020
ERA
2
3
4
5
6
7
8
78
80
82
84
86
88
90
Avg Velo: Off Speed


ERA vs Range of Velo: Off Speed 2020
ERA
2
3
4
5
6
7
8
1
2
3
4
5
Range of Velo: Off Speed


ERA vs Avg Velo Diff: FB and BB 2020
ERA
2 3 4 5 6 7 8
5 10 15 20
Avg Velo Diff: FB and BB


ERA vs Avg Velo Diff: FB and OS 2020
ERA
2 3 4 5 6 7 8
0 2 4 6 8 10 12 14
Avg Velo Diff: FB and OS



The first feature that stands out with all of these graphs is the high degree of randomness. That shows none of these predictors by themselves are a great predictor of ERA. However, any one of them could still play a significant role in predicting ERA when combined with other variables.

The second feature that stands out is the 7 of the 9 best fit lines have negative slope. The first plot with positive slope is the plot with range of velocity for off speed pitches. In this graph in particular, most of the data is between 1-2 mph, with one outlier over 5 mph dominating the fit of the best fit line. The second plot with positive slope is the plot of the difference between average fastball velocity and average off speed pitch velocity. Although this has positive slope, it is very small where it is plausible that no relationship exists between this difference and ERA. Similarly, many of the graphs with negative slope may turn out to not be statistically significant as they appear to be close to 0.

The final feature that stands out is some of these plots have tails that appear to behave differently than the rest of the data. In particular pitchers with fastballs that have an average velocity of +96 mph all have above average ERAs. In the same plot there may also appear to be a pattern for fastballs that average 94 to 96 mph. The range of fastball velocity has a similar tail. 5 of the 6 pitchers with ranges of +3.0 mph have above average ERAs, as well as all five above 3.5 mph. Other tails with above average results are breaking balls that average less than 75 mph and off speed pitches that average above 90 mph.

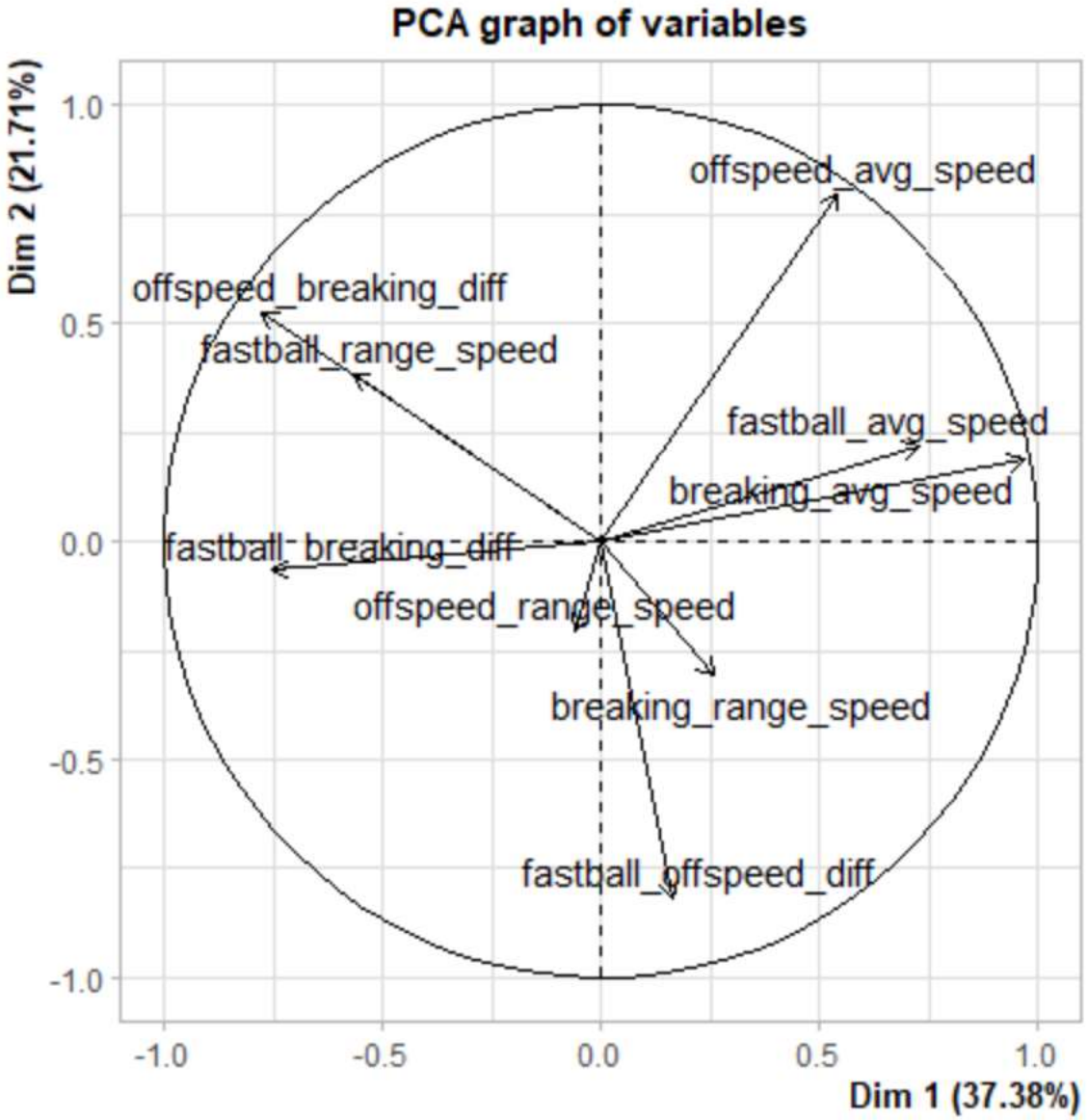
Principal Components Analysis

Correlation Matrix

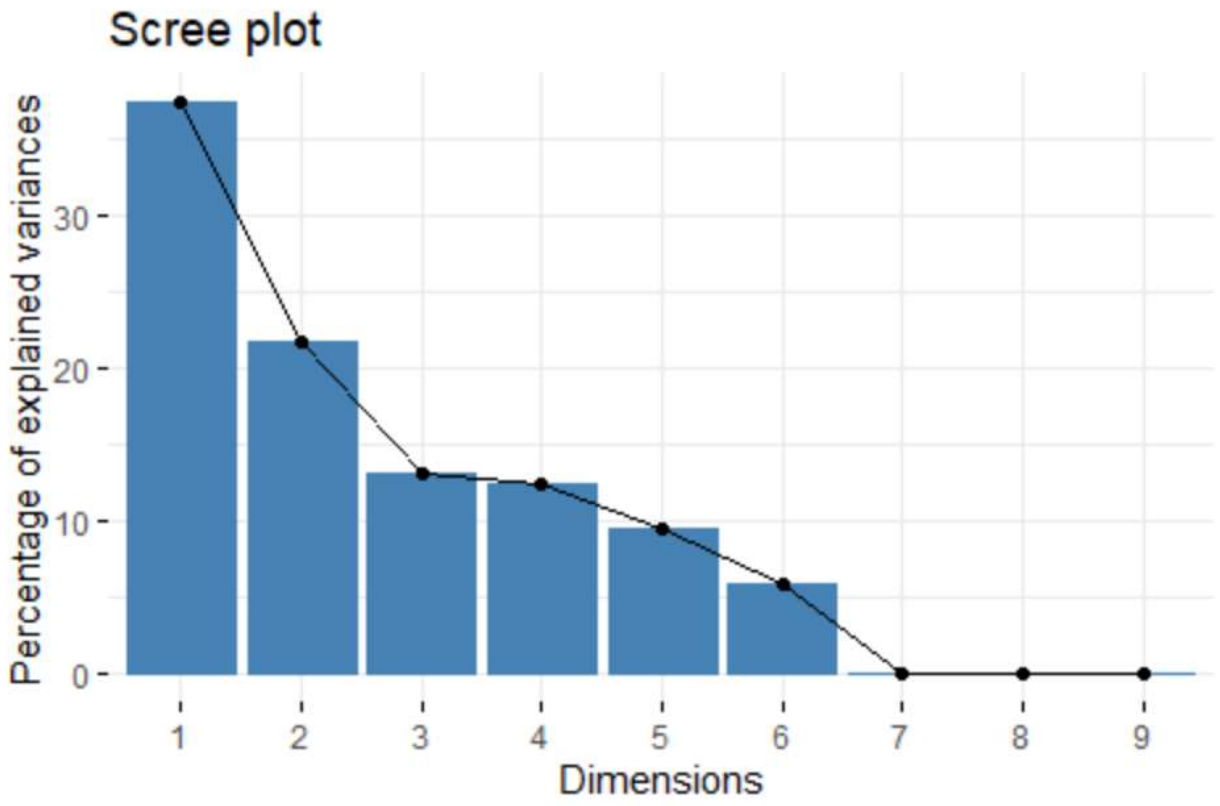
Corr.	FB Avg Speed	FB Range Speed	OS Avg Speed	OS Range Speed	BB Avg Speed	BB Range Speed	FB minus BB	FB minus OS	OS minus BB
FB Avg Speed	1.00	-0.28	0.71	-0.03	0.77	0.06	-0.16	0.27	-0.33
FB Range Speed	-0.28	1.00	-0.03	0.15	-0.39	-0.22	0.32	-0.30	0.50
OS Avg Speed	0.71	-0.03	1.00	-0.10	0.67	-0.02	-0.31	-0.49	0.08
OS Range Speed	-0.03	0.15	-0.10	1.00	-0.04	0.17	0.04	0.10	-0.04
BB Avg Speed	0.77	-0.39	0.67	-0.04	1.00	0.15	-0.76	0.03	-0.69
BB Range Speed	0.06	-0.22	-0.02	0.17	0.15	1.00	-0.17	0.11	-0.23
FB minus BB	-0.16	0.32	-0.31	0.04	-0.76	-0.17	1.00	0.22	0.72
FB minus OS	0.27	-0.30	-0.49	0.10	0.03	0.11	0.22	1.00	-0.52
OS minus BB	-0.33	0.50	0.08	-0.04	-0.69	-0.23	0.72	-0.52	1.00

When looking at the correlation matrix, there are a couple of variables that stand out as being correlated with each other. The first group is the average velocity for fastballs, breaking balls, and off speed pitches. This seems intuitively sensible as most pitchers that can throw one particular pitch fast, likely throw all their different pitches fast. Another set of variables that may have correlation issues is the three average velocity difference variables.

Bi Plot



The bi plot shows all three average velocity variables pointing in the same direction, and each make up a large portion of the first and/or second principal component. Another take away is the three range of velocity variables tend to make up smaller portions of the first and second principal components, especially the ranges for off speed pitches and breaking balls.



The scree plot shows a couple significant features. Although the majority of the variance can be explained by the first two principal components, the third through sixth components also explain a significant portion of the variance. However, the most notable observation is the seventh, eighth, and ninth components make up none of the variance. This shows that three of the variables being considered may actually not be necessary to be included in analysis.

Generalized Linear Models

Generalized linear models were fit using a top down approach and log transformation. Variables were eliminated through performing t-tests, and AIC was used to help determine an optimal model. The first model uses an intercept and all 9 predictor variables, with each variable being standardized prior to fitting. After an optimal model was determined based on that initial model, further testing was done to see if changing certain predictor variables to be indicator variables provided a better fit than their numerical values. This was inspired by the observation that several of the tails in the scatter plots behave different than the rest of the plot.

The overarching purpose of this exercise is not necessarily to come up with a perfect model to predict a pitchers future ERA. Rather, it is to continue to learn more about the relationship between our predictor variables and ERA. In particular, the main benefit to using a multilevel model is to see if any masked or spurious relationships exist which we were not able to see in our scatter plots.

Due to brevity, not all models that were fit will be shown. Instead, only the interesting/significant models will be displayed.

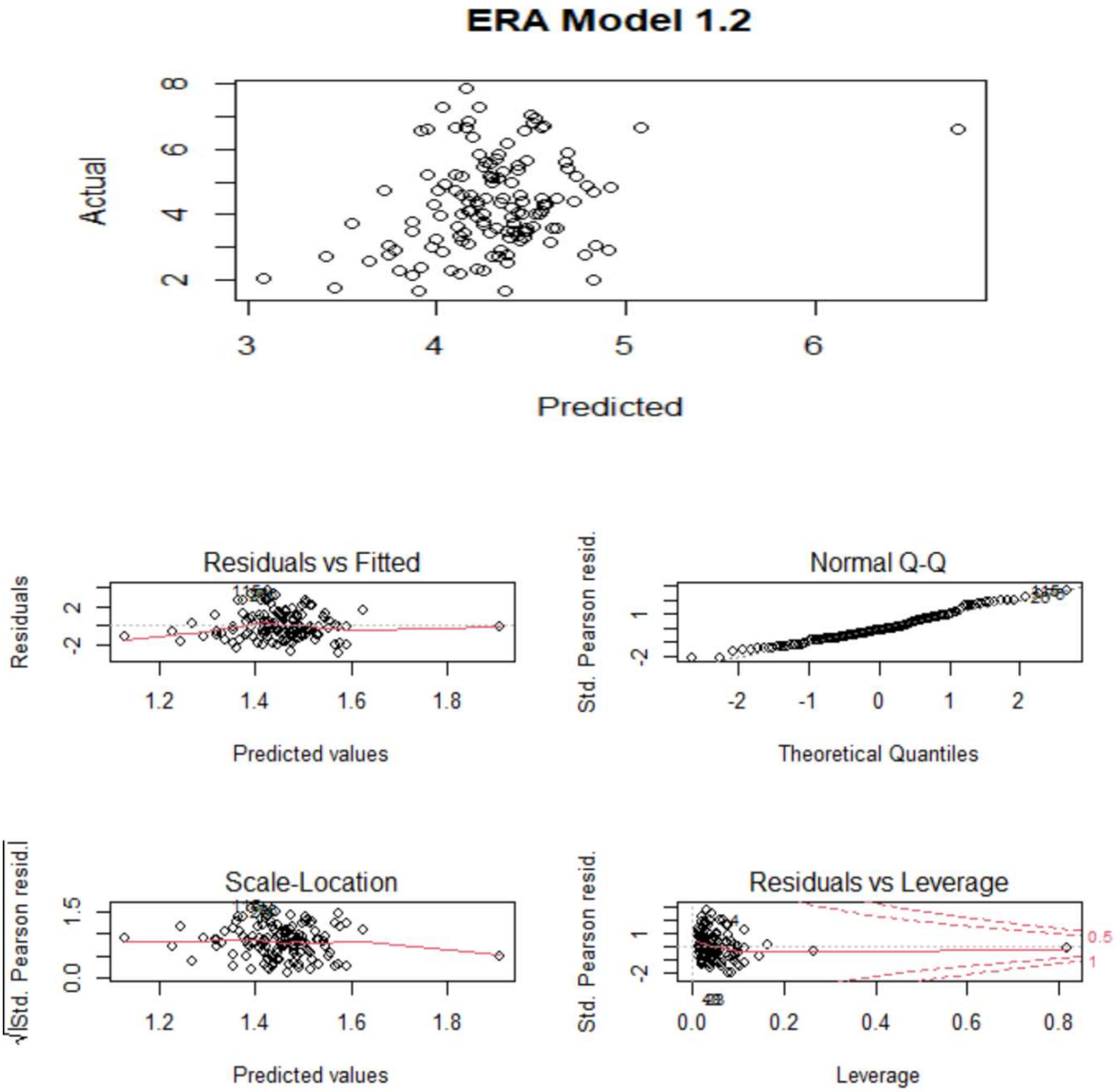
Model 1.1:

```
Coefficients: (3 not defined because of singularities)
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                1.454533   0.028692  50.695   <2e-16
fastball_avg_speed_std    -0.049397   0.048913  -1.010   0.3145
fastball_range_speed_std  -0.071384   0.037205  -1.919   0.0573
breaking_avg_speed_std     0.006733   0.050915   0.132   0.8950
breaking_range_speed_std  -0.032788   0.029783  -1.101   0.2731
offspeed_avg_speed_std    -0.014476   0.045645  -0.317   0.7517
offspeed_range_speed_std   0.043786   0.021800   2.009   0.0468
fastball_breaking_diff_std       NA         NA      NA       NA
fastball_offspeed_diff_std       NA         NA      NA       NA
offspeed_breaking_diff_std       NA         NA      NA       NA
```

The very first model tested produced interesting results with the value of the intercepts for the three difference in average velocity variables all being NAs. The reason this occurred is due to a multicollinearity issue, since each of these variables are just linear combinations of other variables in the model. This explains why the scree plot showed all the variance can be explained in the first 6 principal components. Moving forward, the next model will drop all three of these variables.

Model 1.2:

```
Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                1.454533   0.028692  50.695   <2e-16
fastball_avg_speed_std    -0.049397   0.048913  -1.010   0.3145
fastball_range_speed_std  -0.071384   0.037205  -1.919   0.0573
breaking_avg_speed_std     0.006733   0.050915   0.132   0.8950
breaking_range_speed_std  -0.032788   0.029783  -1.101   0.2731
offspeed_avg_speed_std    -0.014476   0.045645  -0.317   0.7517
offspeed_range_speed_std   0.043786   0.021800   2.009   0.0468
```



The second model has eliminated the multicollinearity issue, and can serve as a starting point. Initially it looks like two predictor variables, fastball and off speed ranges, may be significant by having low p-values. However, as variables are eliminated one at a time, other variables may become significant with more testing.

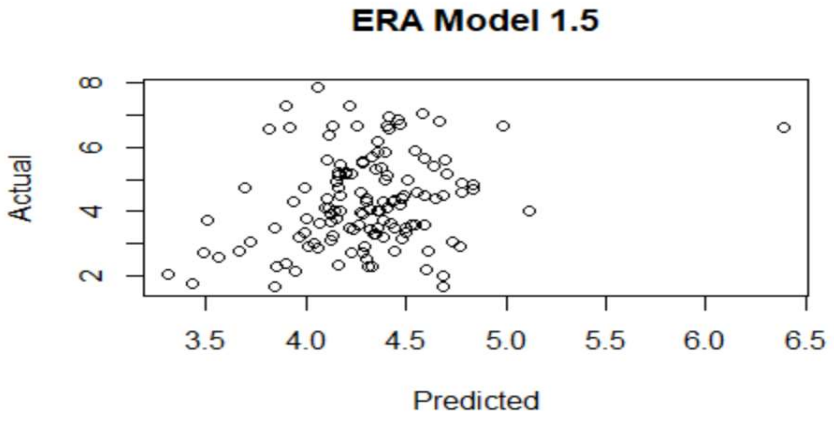
By looking at the scatter plot of predicted vs actual, this model does not necessarily give the best predictions particularly in the middle. However, it does show a general shaped of positive slop which is good, and in general the left side of the graph seems to have predictions that are closer to the actual results. Going by the old saying, "no model is perfect, but some are useful," this may be a sign of how this model is useful. Not all predictions are necessarily all that great, but if a pitcher is predicted to have a low ERA, he likely will be relatively close to that prediction, which could be very useful information.

When testing the residuals to make sure a linear model is the best fit, a few things stand out. First, the residual plot shows a lot of randomness in the middle, which is good. However, the ends have residuals that systematically get closer, suggesting this model may not be the best fit and could be better if it had a transformation. Similarly the q-q plot is a fairly straight line in the middle, but peels off slightly on the ends. For the purposes of this study, this fact was ignored and all other models were test under the same fit as it should still suffice to help better understand the predictor variables and their relationship to the response.

Overall the AIC of this model was 468.1. The next several models tested dropped one insignificant variable at a time.

Model 1.5:

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                1.45475    0.02849  51.056   <2e-16 ***
fastball_avg_speed_std    -0.05692    0.02919  -1.950   0.0534 .
fastball_range_speed_std -0.06850    0.03306  -2.072   0.0403 *
offspeed_range_speed_std   0.04112    0.02170   1.895   0.0604 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Model 1.5 is the first model that has only significant predictors. Two of the predictors are the same as what appeared to be significant in Model 1.2, fastball and off speed range speed. However, fastball average velocity which first seemed insignificant has become significant after dropping off speed and breaking ball average velocity.

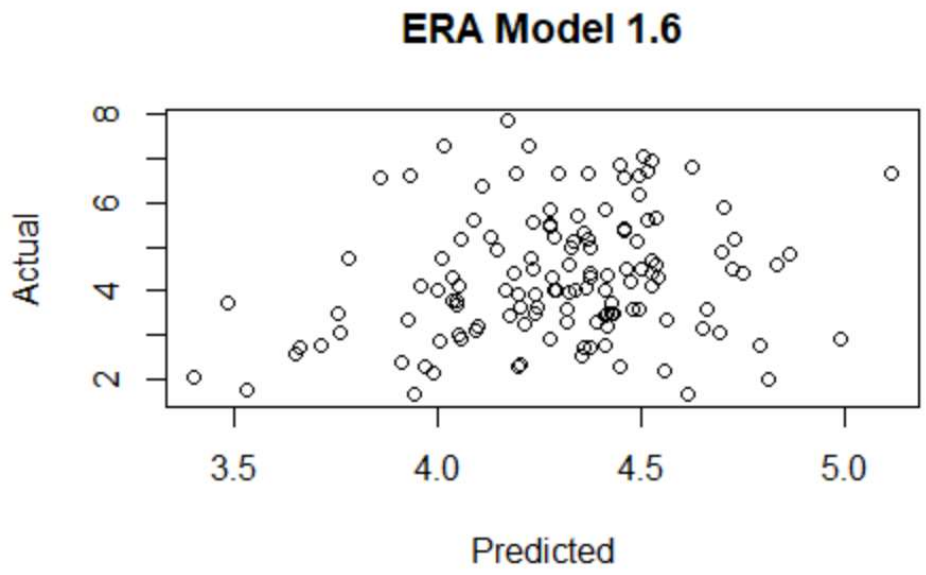
This suggests that a spurious relationship exists between average fastball velocity, average off speed and breaking ball velocity, and ERA. What that means is individually each predictor variable (the three average velocities) has a relationship with the response variable (ERA). However, because the predictor variables are highly correlated, as we saw in PCA, when all the variables are used in the same model they no longer all still have a relationship with the response variable. In fact there is one predictor variable, in this case average fastball velocity, which is seen as having a predictive relationship with the response variable. The other variables, although they may be correlated with the response, do not necessarily have a causal relationship with the response. In terms of this study, throwing a hard breaking ball or off speed pitch does not necessarily suggest a pitcher will have a lower ERA. However, if a pitcher throws a fast breaking ball and off speed pitch, that same pitcher likely also throws a fast fastball which does suggest he is more likely to have a lower ERA.

There is one more observation that stands out in this model. Although all variables are significant, average fastball velocity and range of fastball velocity both have negative relationships with ERA. This means as a pitcher throws harder, or as a pitcher is able to better vary the speed of his fastball, the lower is ERA is expected to be. This makes intuitive sense. However, the range of off speed pitch velocity has a positive relationship with ERA. This suggests as a pitcher increases the difference in speed of his off speed pitch, the higher his ERA is likely to be. That does not seem to make intuitive sense. Looking back at the scatter plot for range of off speed velocity vs ERA, there is a seemingly positive relationship that is predominately driven by one outlier with a range greater than 5. With that note, although this appears to be statistically significant in our model, is may not actually be representative of future data, and may be best ignored.

Overall the AIC of this model is 463.58. The next two models will continue eliminating one variable at a time. After that, the next models will go back to model 1.2, remove the range of off speed velocity and continue the practice of dropping one variable at a time to see if not including the range of off speed velocity makes another variable significant that otherwise wouldn't be.

Model 1.6:

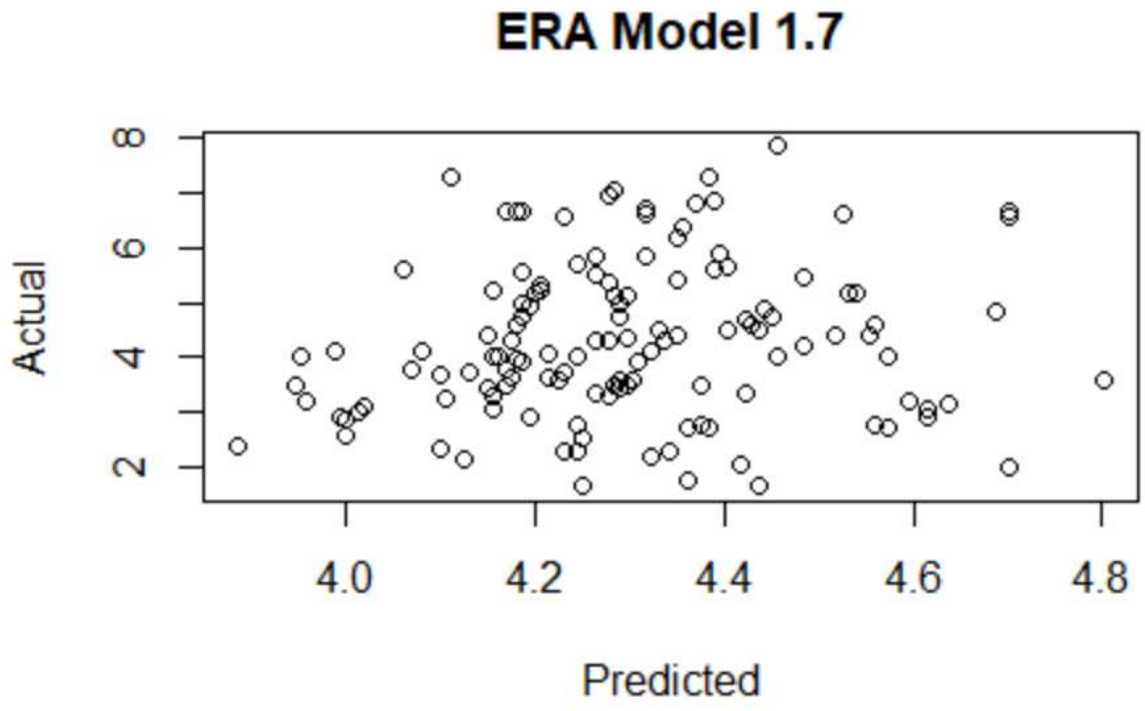
```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            1.45577    0.02862  50.863   <2e-16 ***
fastball_avg_speed_std -0.05644    0.02933  -1.924   0.0566 .
fastball_range_speed_std -0.06018    0.03243  -1.856   0.0658 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



The AIC of this model is 464.45.

Model 1.7:

```
Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.45720    0.02883  50.553   <2e-16 ***
fastball_avg_speed_std -0.04105    0.02859  -1.436    0.154
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



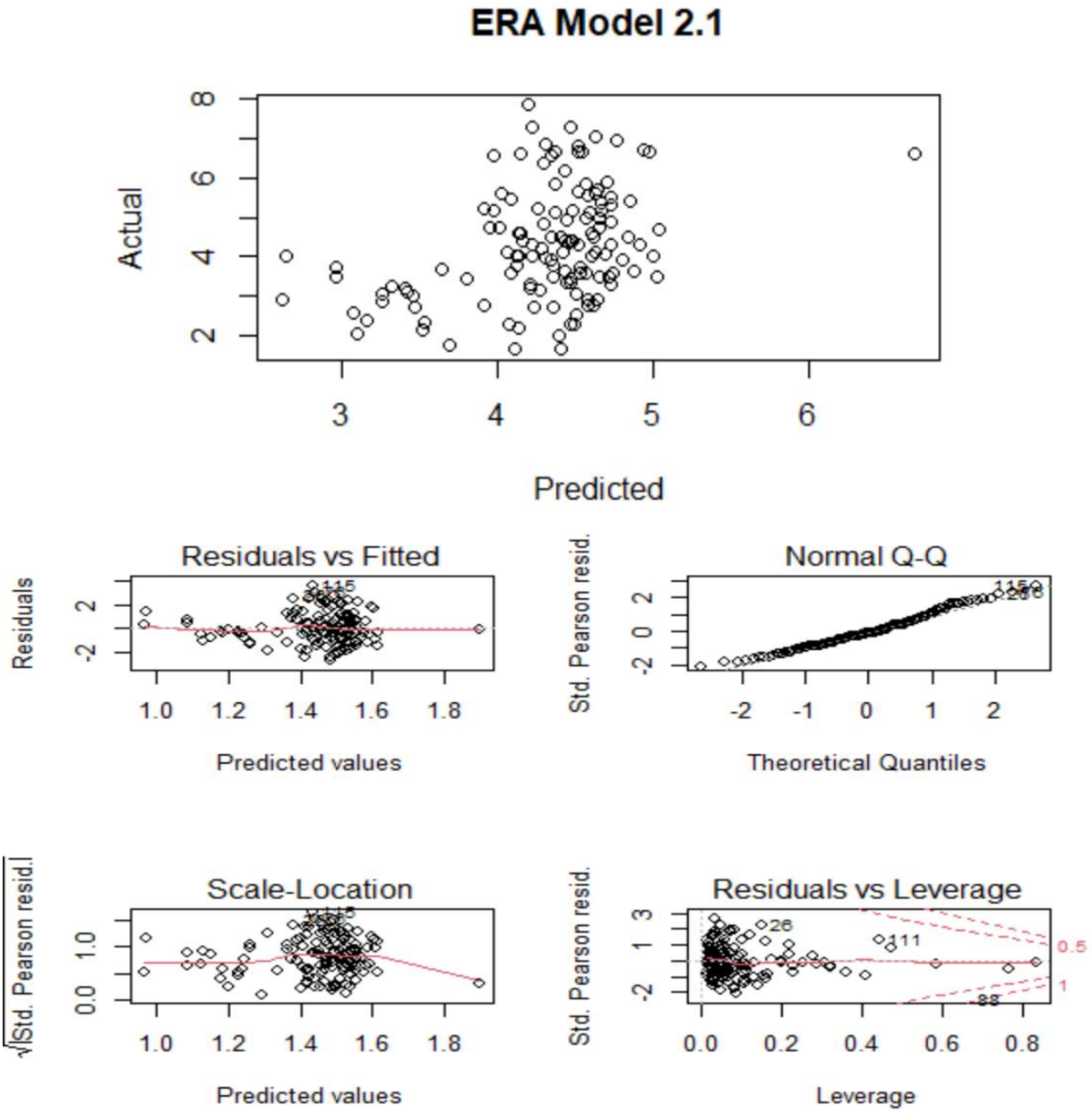
Overall, the AIC of this model is 466.56. Although Model 1.5 has the lowest AIC, suggesting the best fit without overfitting, due to the outlier situation, Model 1.6 will be chosen as the optimal model thus far with the second lowest AIC. After removing the range of off speed velocity from Model 1.2, no additional relationships were discovered, thus Model 1.6 remains the optimal model.

Again, the purpose of this study is not to necessarily create the best model to predict a pitcher's ERA. If that were the case, a lot more variable would be considered. Instead the purpose is to try to see what can be understood about the relationship between average velocity and range of velocity and a pitcher's ERA. So far it appears that to predict a pitcher's success based on a radar gun only, only that pitcher's fastball is important. Within that pitch though just as important as how hard a pitcher can throw his fastball is how well he can vary the speed of his fastball.

Going forward, the next steps will be to test more models to see if switching from numerical values to indicator variables produces different results. The first set will again be a top down structure, this time adding an indicator variable that equals 1 if the average fastball velocity is greater than 94 and 0 otherwise. Cross effects will also be added for all variables.

Model 2.1:

```
                                          t value Pr(>|t|)
(Intercept)                                45.933   <2e-16 ***
fastball_avg_speed_std                      0.717   0.4751
fastball_range_speed_std                   -1.531   0.1285
breaking_avg_speed_std                      0.019   0.9848
breaking_range_speed_std                   -1.076   0.2840
offspeed_avg_speed_std                     -0.806   0.4217
offspeed_range_speed_std                    1.980   0.0501 .
fastball94plus                              0.354   0.7243
fastball_avg_speed_std:fastball94plus      -1.583   0.1162
fastball_range_speed_std:fastball94plus    -0.975   0.3314
breaking_avg_speed_std:fastball94plus       0.223   0.8242
breaking_range_speed_std:fastball94plus    -0.926   0.3564
offspeed_avg_speed_std:fastball94plus       1.083   0.2808
offspeed_range_speed_std:fastball94plus     0.598   0.5509
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



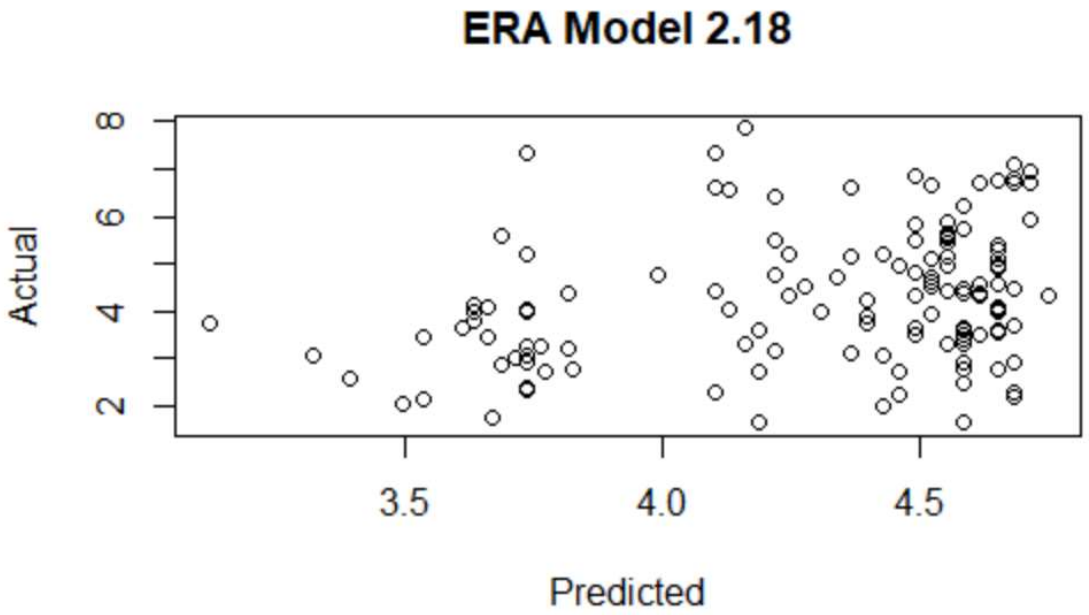
The initial look at the full model with the +94 mph indicator for fastball velocity only shows one significant variable, range of off speed velocity. That variable will again be removed due to having a positive effect because of the outlier. Going from there variable will continued to be removed one at a time until only significant effects remain to obtain the optimal model.

An interesting note of this model is the backwards L shape. This model doesn't appear to be a great linear predictor. However, the 20 best predictors were all above average, after that the results were pretty random. Again this may not be the best in terms of a linear relationship, but may provide helpful in finding above average pitchers with little error. Similar to Model 1.2, the residual plot would suggest a transformation may be helpful here, but again since the focus isn't necessarily to build an accurate model, but understand the relationship, the rest of the models will not include any transformations.

This model has an AIC of 472.57. It makes sense that it is higher than the AICs seen in the other models thus far, as adding all the cross effects has doubled the number of parameters.

Model 2.18:

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                1.49689    0.03006  49.791  < 2e-16
fastball_range_speed_std  -0.05465    0.03068  -1.781  0.07725
fastball94plus            -0.21806    0.08020  -2.719  0.00746
```

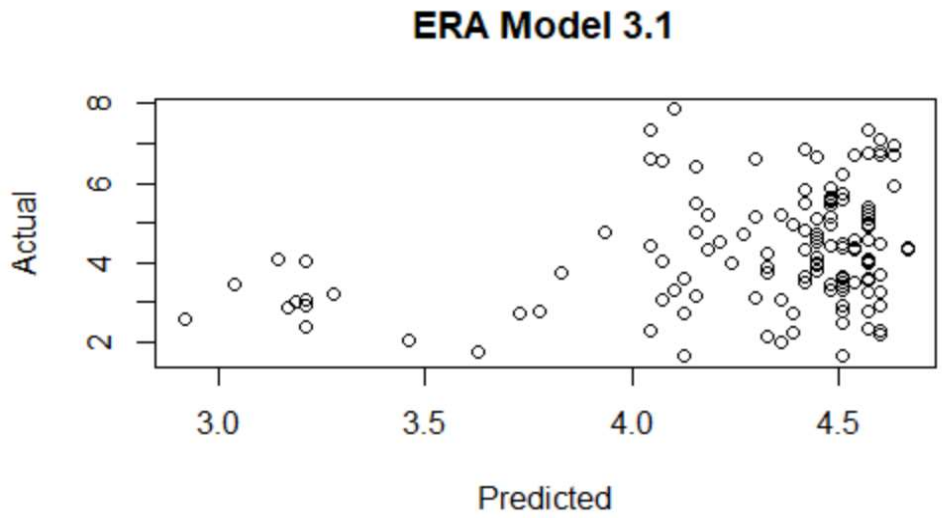


After removing all insignificant variables, in turns out there were no cross effects that remain, and the best model ended up being very similar to Model 1.6. The one difference is switching average fastball velocity to the indicator variable of +94 mph.

The AIC of this model is 460.03. The next set of models will be comparing the 94+ mph indicator to a +96 mph indicator, and adding a 94 to 96 mph indicator as well.

Model 3.1:

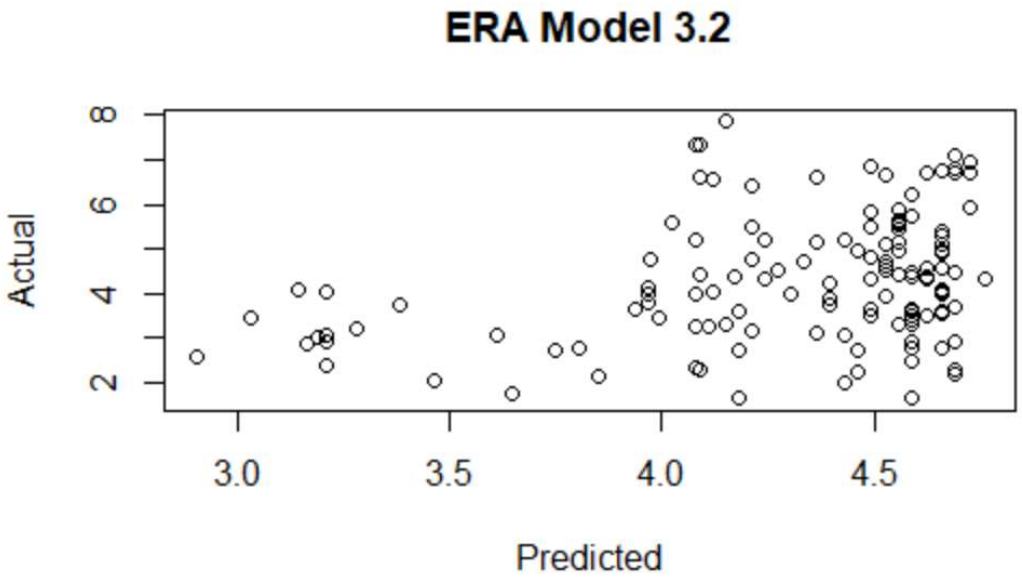
```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                1.48016    0.02847  51.985   <2e-16
fastball_range_speed_std  -0.05334    0.03051  -1.748   0.0828
fastball96plus            -0.35311    0.14094  -2.505   0.0135
```



The AIC of this model is 459.96.

Model 3.2:

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                1.49681    0.02991  50.048  < 2e-16
fastball_range_speed_std  -0.05660    0.03052  -1.854  0.06601
fastball94to96            -0.13194    0.09147  -1.443  0.15162
fastball96plus            -0.37125    0.14064  -2.640  0.00934
```



This model has an AIC of 459.61.

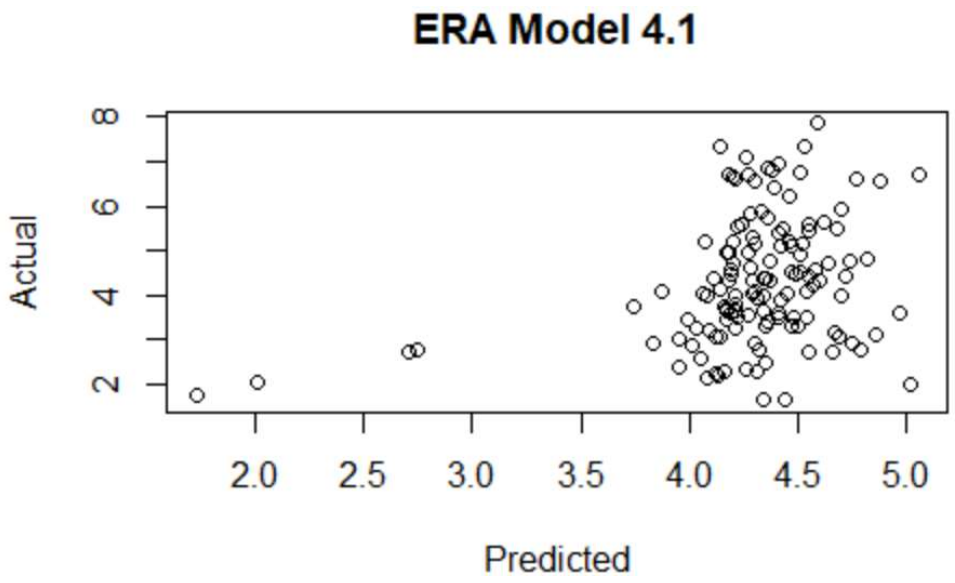
In moving from Model 2.18 to 3.1, and then again to 3.2, the AIC decreased marginally both times. The backwards L shape returned in Model 2.18, and became more excessive in Model 3.1 and 3.2. One interesting aspect is the 94 to 96 mph indicator variable itself is not significant, but the addition of that variable makes the +96 mph and the fastball range velocity

more significant to the point the AIC overall decreased. As for moving forwards, all of these models as well as Model 1.6 will be used with test data for comparison sake.

The next set of models will be testing if an intercept should be used for the range of fastball velocity. While testing this, the standardized value for average fastball velocity will be used instead of the fastball intercepts.

Model 4.1:

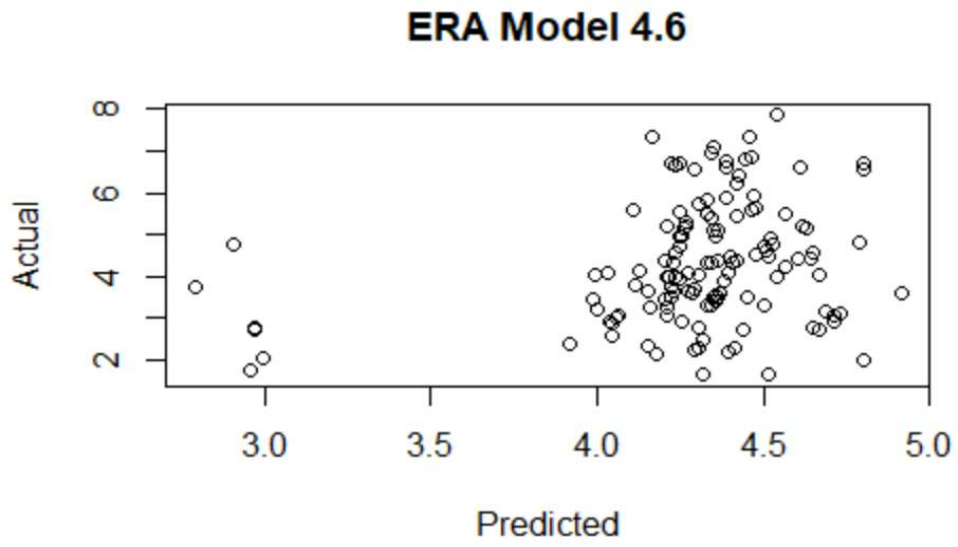
```
Coefficients:
                                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                                    1.4709634  0.0297952  49.369   <2e-16
fastball_avg_speed_std                        -0.0488542  0.0494554  -0.988    0.325
fastball_range_speed_std                      -0.0005058  0.0462687  -0.011    0.991
breaking_avg_speed_std                         0.0295633  0.0525275   0.563    0.575
breaking_range_speed_std                      -0.0196477  0.0302760  -0.649    0.518
offspeed_avg_speed_std                        -0.0261255  0.0470905  -0.555    0.580
fastball_range3.0plus                          1.7087033  1.8081311   0.945    0.347
fastball_avg_speed_std:fastball_range3.0plus   0.0259174  0.7320741   0.035    0.972
fastball_range_speed_std:fastball_range3.0plus -0.6654885  0.6385946  -1.042    0.299
breaking_avg_speed_std:fastball_range3.0plus   0.0678745  0.8515623   0.080    0.937
breaking_range_speed_std:fastball_range3.0plus 0.3051225  0.7627610   0.400    0.690
offspeed_avg_speed_std:fastball_range3.0plus   0.1916983  0.9241120   0.207    0.836
```



Similar to the other models, there isn't too much going on in the full model here. The predicted vs actual scatter plot is concerning in having a few predictions significantly further from the rest, which may lead to overfitting. The AIC of this model is 475.56.

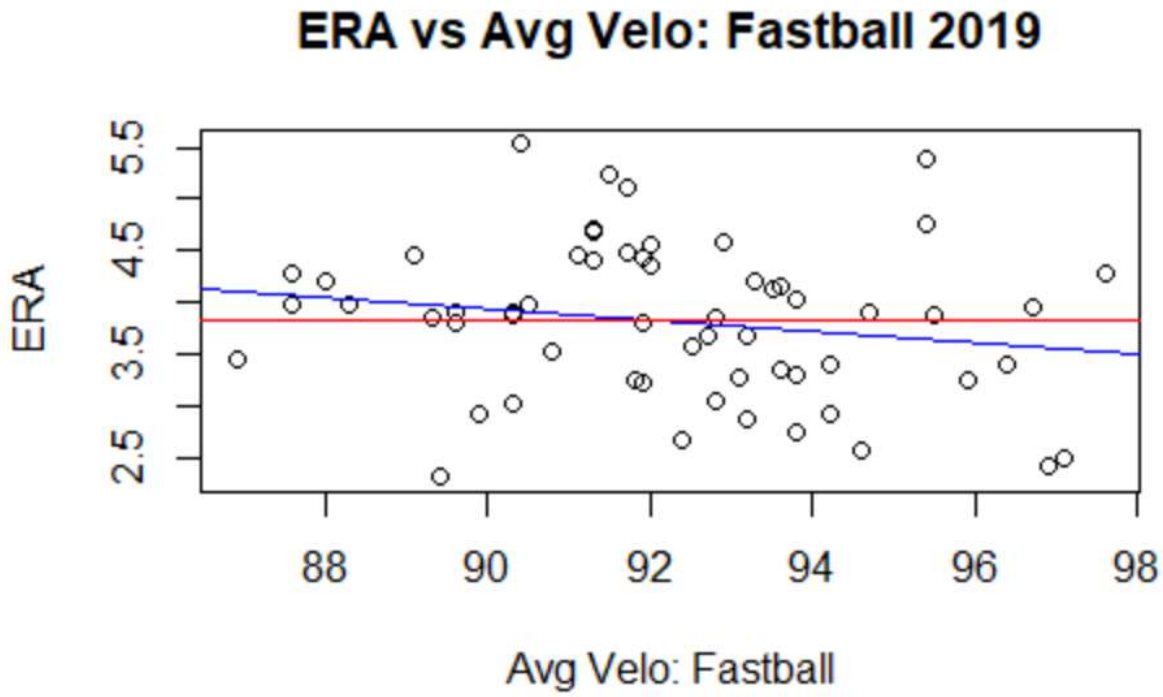
Model 4.6:

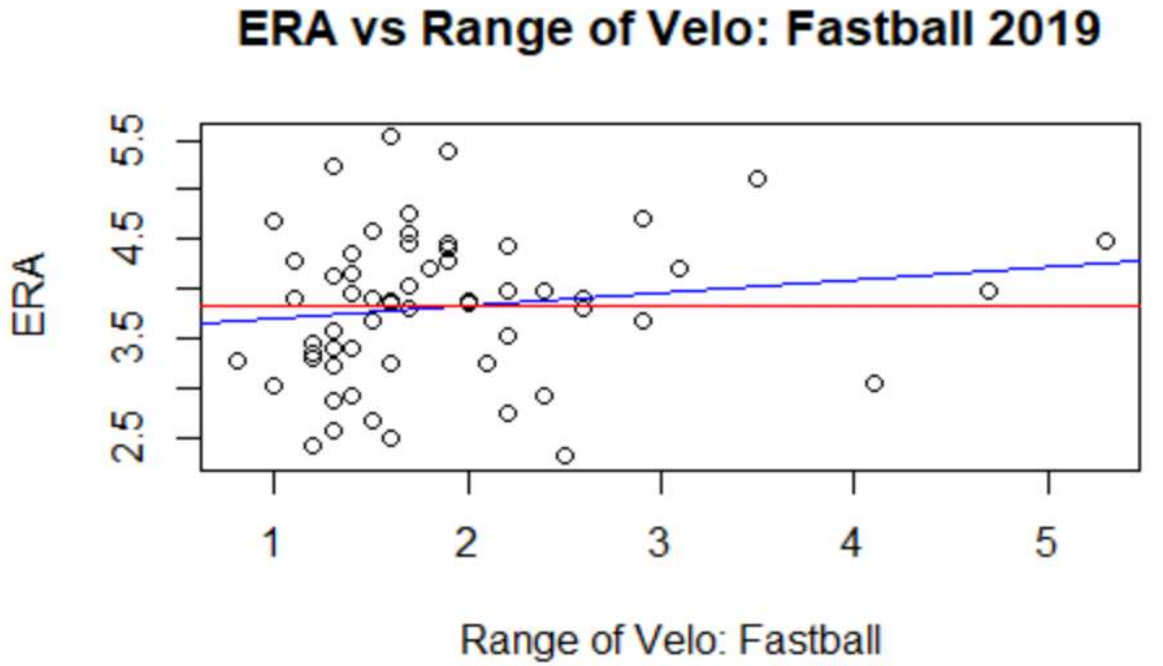
```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.47255    0.02847  51.727   <2e-16 ***
fastball_avg_speed_std -0.04398    0.02767  -1.589   0.1144
fastball_range3.0plus  -0.40589    0.19513  -2.080   0.0395 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

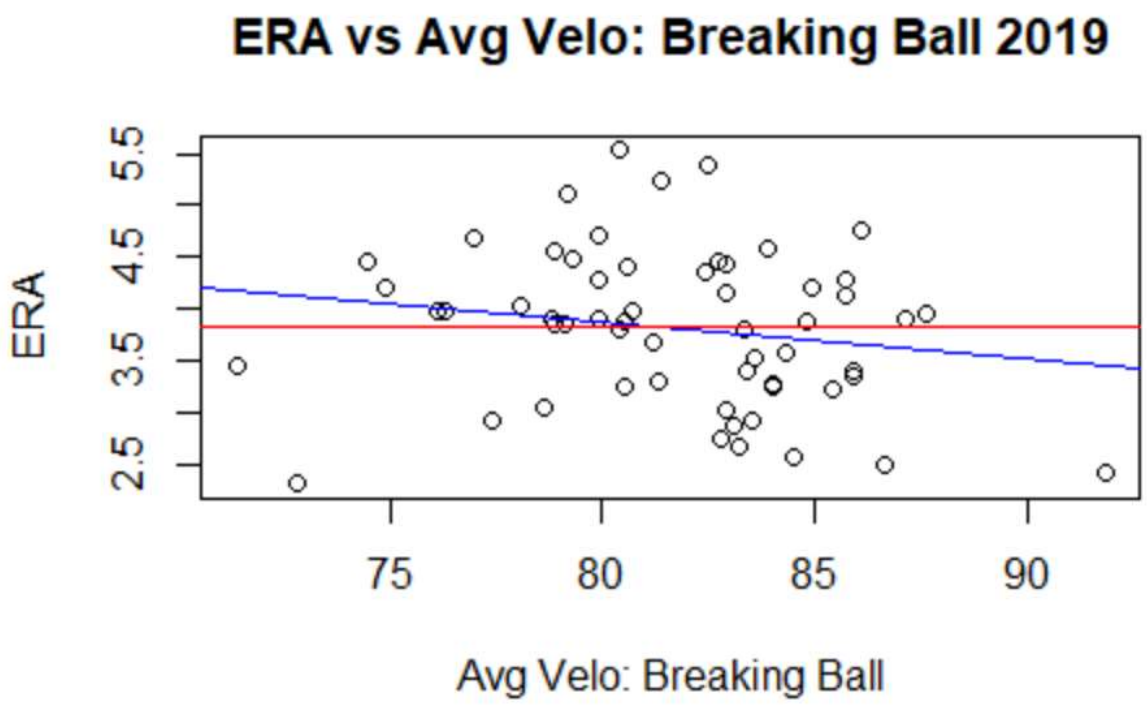


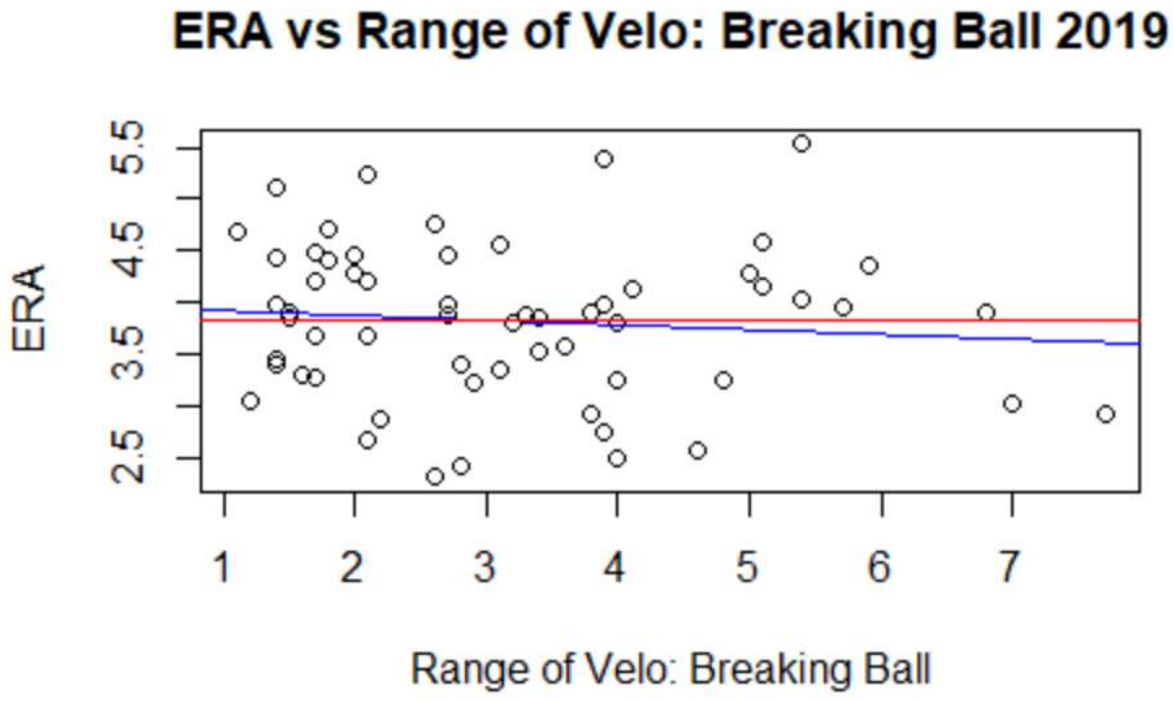
After removing variables, eventually all variables are removed except the indicator variable and fastball average speed. In this model fastball average speed is insignificant, and the AIC is 462.17, which is higher than the previous models. This model is not optimal.

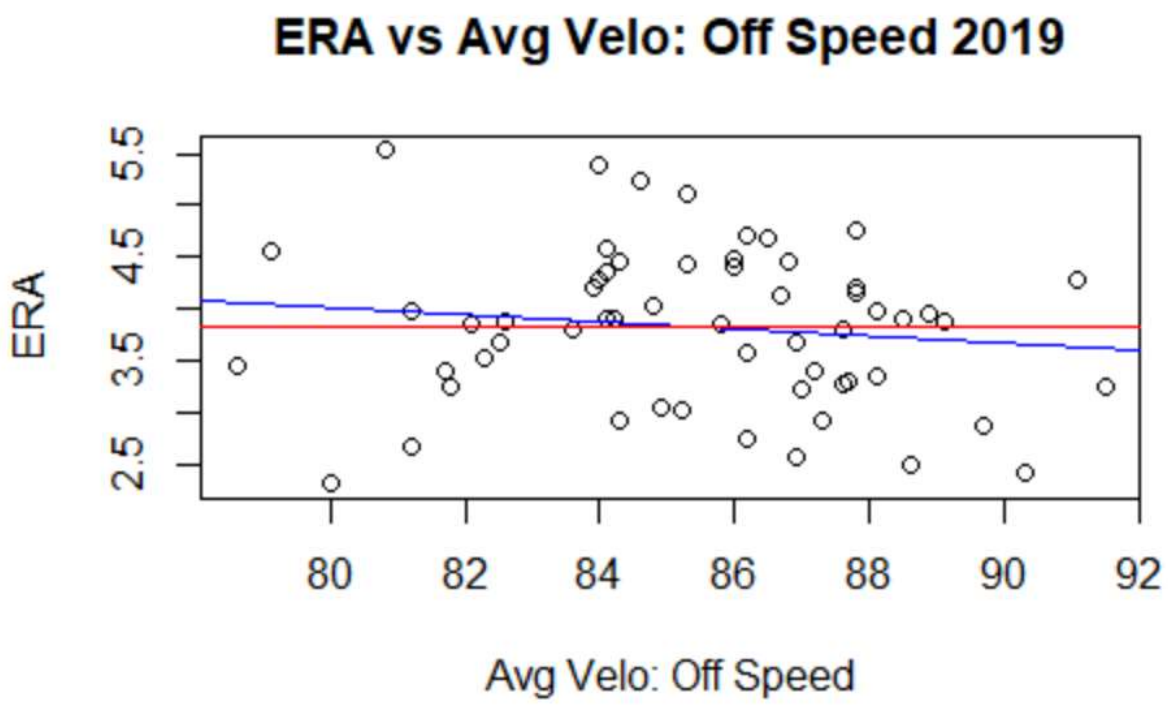
Comparing to Test Data

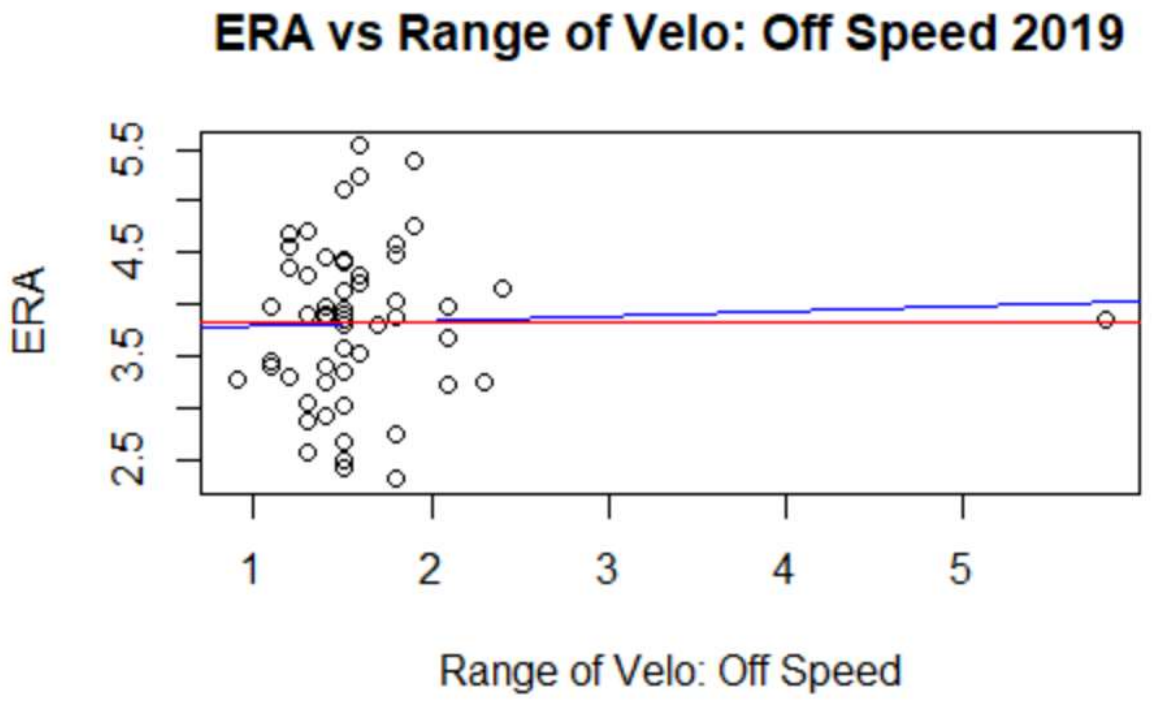


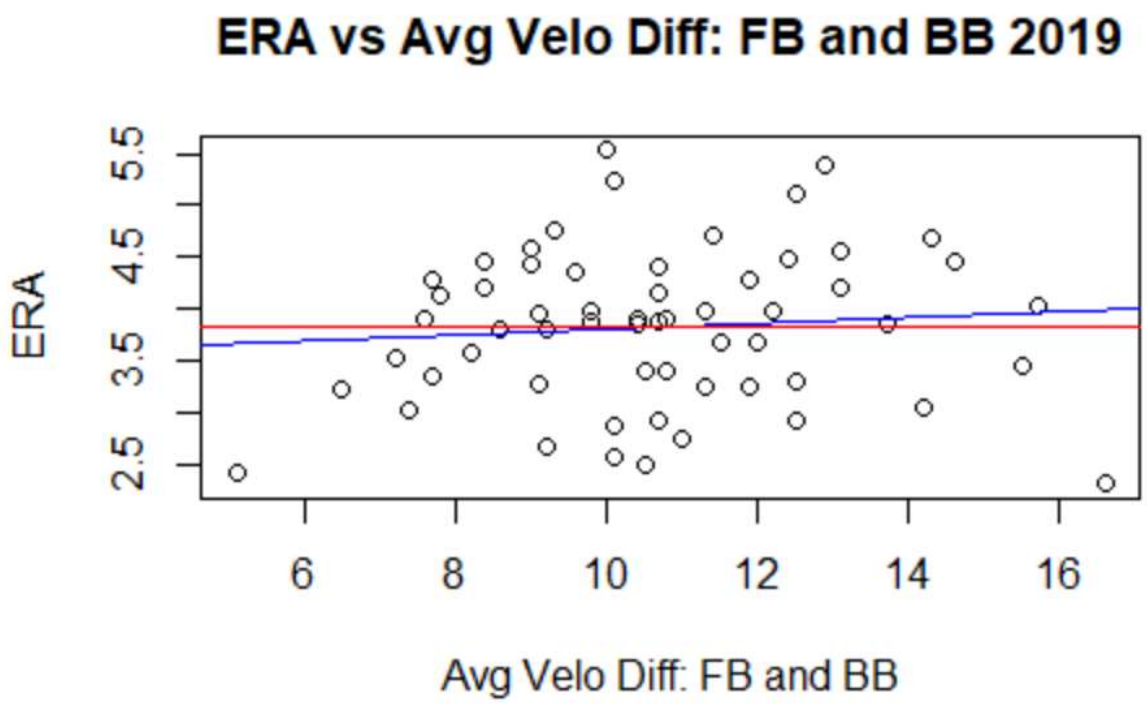


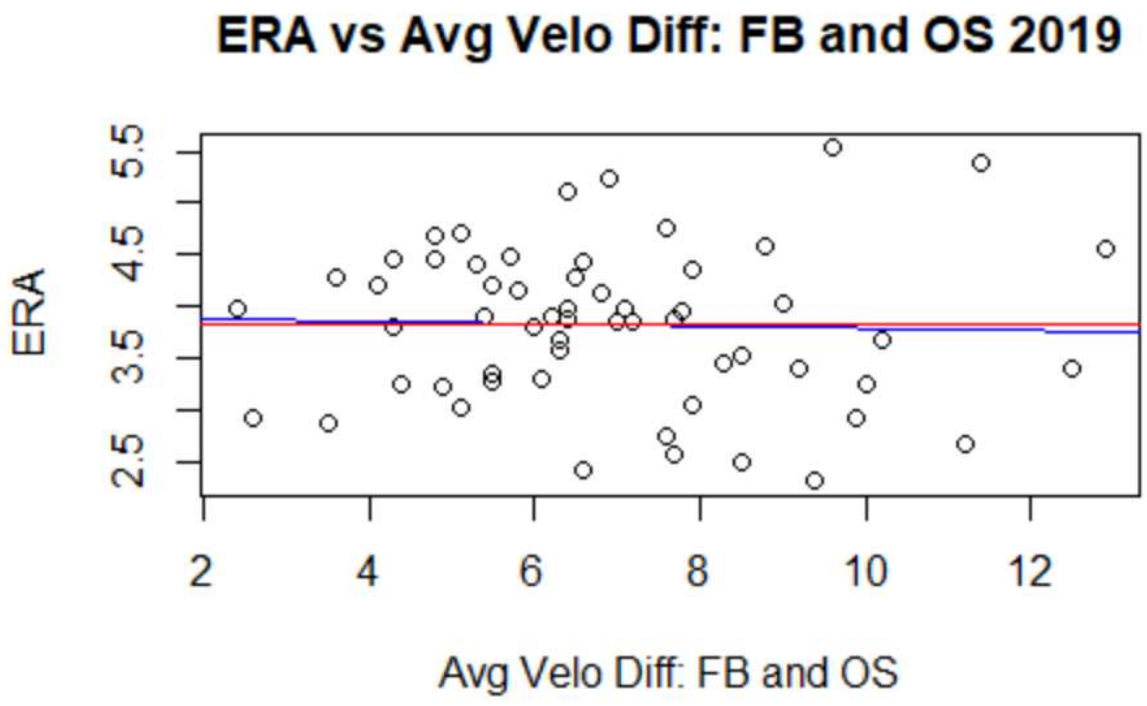

ERA vs Avg Velo: Breaking Ball 2019
ERA
2.5
3.5
4.5
5.5
75
80
85
90
Avg Velo: Breaking Ball

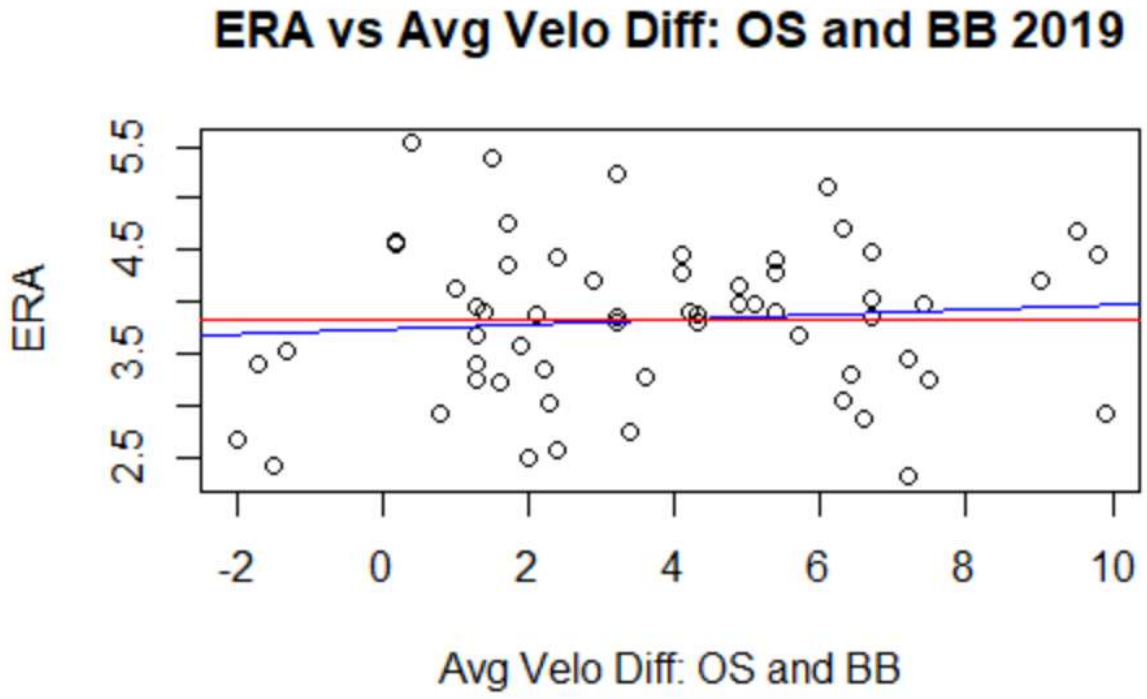

ERA vs Range of Velo: Breaking Ball 2019
ERA
2.5
3.5
4.5
5.5
1
2
3
4
5
6
7
Range of Velo: Breaking Ball


ERA vs Avg Velo: Off Speed 2019
ERA
2.5
3.5
4.5
5.5
80
82
84
86
88
90
92
Avg Velo: Off Speed

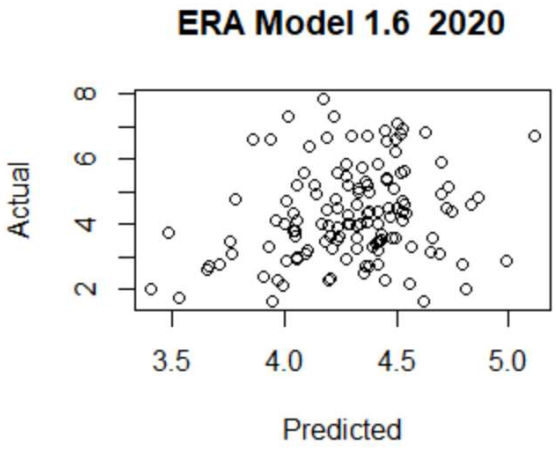

ERA vs Range of Velo: Off Speed 2019
ERA
2.5
3.5
4.5
5.5
1
2
3
4
5
Range of Velo: Off Speed


ERA vs Avg Velo Diff: FB and BB 2019
ERA
2.5
3.5
4.5
5.5
6
8
10
12
14
16
Avg Velo Diff: FB and BB

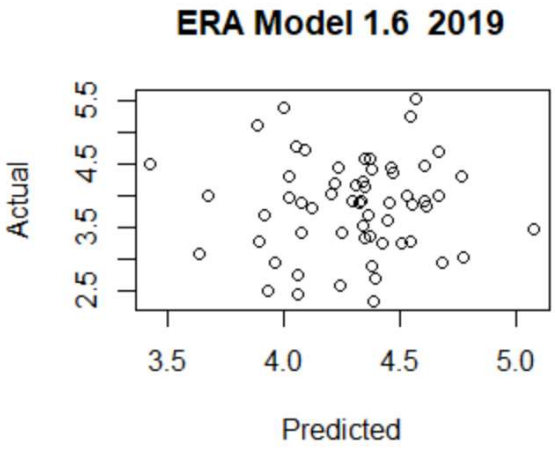

ERA vs Avg Velo Diff: FB and OS 2019
ERA
2.5
3.5
4.5
5.5
2
4
6
8
10
12
Avg Velo Diff: FB and OS



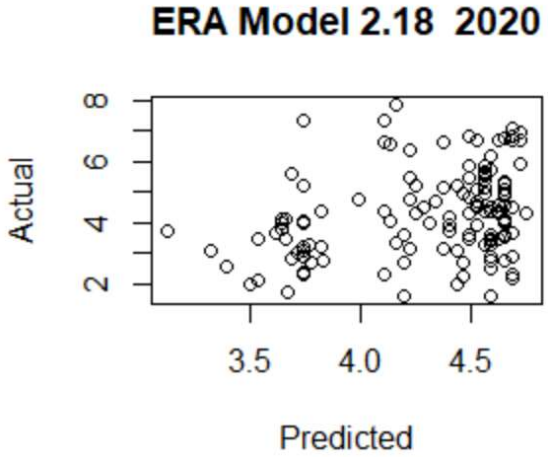
Data from the 2019 season is being used as test data for this study. The first comparison is in the scatterplots. One thing of note here is the average fastball, breaking ball, and off speed velocity all have a negative relationship to ERA, which is consistent with the training data, or 2020. The range variables have some noticeable differences between the two years. In particular, the range of fastball velocity flipped from having a negative slope to a positive slope. Another significant feature was in 2020, 5 of the 6 pitchers with ranges above 3 mph were above average. In 2019, 4 of the 5 pitchers with ranges above 3 were below average. The range of breaking ball velocity stayed fairly consistent as before, having a slightly negative but close to 0 slope. The range of velocity for off speed pitches continued to have a positive slope, but it changed to be slightly positive and very close to 0. The slopes for the three variables of the differences between fastball, breaking ball, and off speed average speed all remained close to 0.



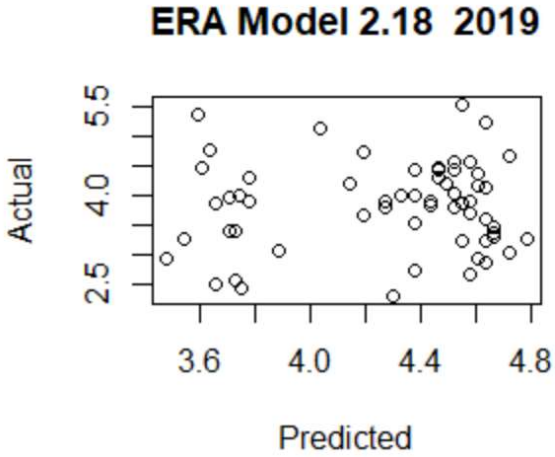
MSE = 1.909



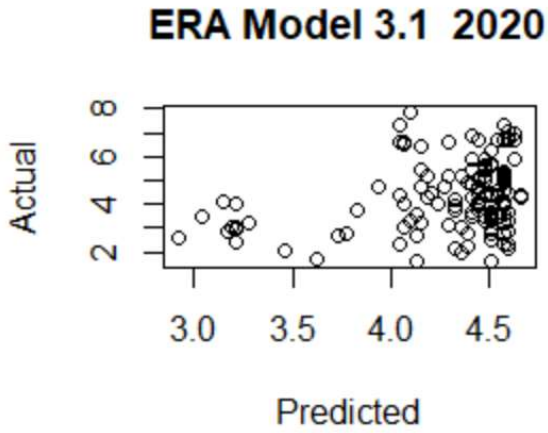
MSE = 0.841



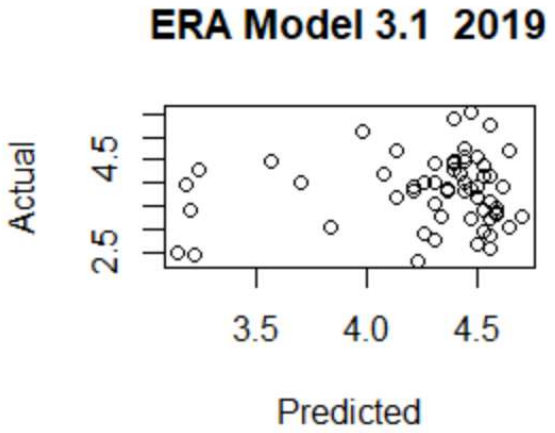
MSE = 1.846



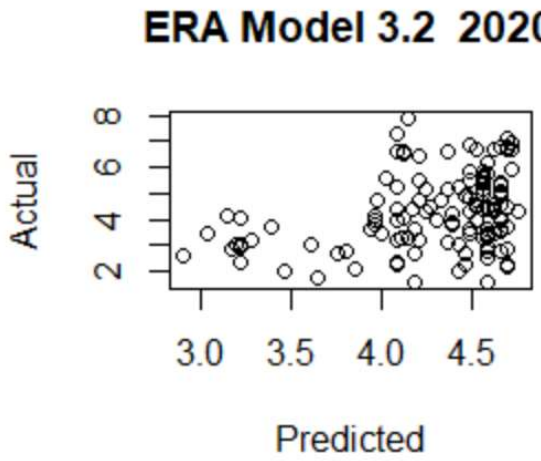
MSE = 0.854



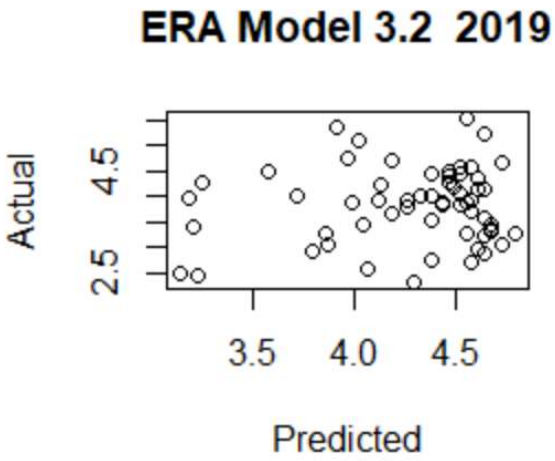
MSE = 1.845



MSE = 0.840



MSE = 1.812



MSE = 0.850

When comparing the results of the models on 2020 and 2019 data side by side there are a couple observations of note. The first note is none of these models appear to have the best predictive accuracy. This is shown by significant randomness not only in the test data, but even the training data. Another significant note is the range of actual results is significantly smaller in 2019 than in 2020. This is likely for two main reasons, both stemming from 2019 being a full season, and 2020 being shortened. With a longer season, outlier 'rough starts' where a pitcher uncharacteristically gives up a large number of runs has more time to be weighted out with other starts. In a shortened season these starts can carry more weight. Additionally, in a short season a team might be more willing to continue to let a pitcher play if he is performing poorly with the hope of he is going through a rough stretch and will get better throughout the year. The longer the season lasts, the more time there is for a team to give up on a player performing poorly which would eliminate that pitcher from the dataset if he doesn't throw enough innings by the end of the year. Essentially if a pitcher has an ERA above 6, he needs to either perform better or he won't keep pitching. The result of this shrinks the range of actual ERA, and decreases the MSE by bringing more observations closer to average.

Model 1.6 appears to poorly fit the training data producing mostly random results, but does show a slight positive slope suggesting the higher the prediction is the higher the result is. The test data is similar with a slight positive slope, but still mostly random results.

Model 2.18 appears to be an improvement based on training data. This was seen earlier due to a smaller AIC, we continue to see this with a decrease in MSE. In addition this model produced a backwards L scatterplot. However, when comparing to the test data, the shape of the scatterplot transforms becoming random. Additionally, the MSE actually increased in the test data. This suggest that Model 2.18 may actually be over fitting the data.

Model 3.1 also appears to be an improvement over model Model 1.6. Similar to Model 2.18, Model 3.1 has a smaller MSE, and an even more exaggerated backwards L shape. Unlike Model 2.18, the test data in Model 3.1 also shows an improvement in MSE from Model 1.6. Additionally the backwards L shape remained present in the test data, although less visible due to

the decrease in max ERA. Again even though this model does not have the best linear shape, it is intriguing that in both the training and test data the best 10-15% of observations based on predictions all had an actual ERA of about 4.5 or less.

When comparing Model 3.2 to Model 3.1, it appears there is an improvement as noted by the decrease in MSE. It should be noted though, that this decrease came at the cost of adding a parameter. When evaluating the test data, the MSE actually rose slightly. This again could be a sign of overfitting.

Again none of these models are perfect, or even great models. However, of the models, Model 3.1 would be the best model, using the range of fastball velocity and an indicator variable of an average fastball velocity of +96 mph.

Conclusion

So what does all this mean? Is it more important for a pitcher to throw fast, or does his ability to change speeds matter just as much if not more? In sum, I would say how hard a pitcher throws is more important than his ability to change speeds. In particular, it is best if a pitcher can maintain an average fastball velocity above 96 mph. However, it should be made clear, that a pitchers ability to change speeds and willingness to do so are two separate things. In this study, the velocity difference within the same type of pitch and amongst different pitch types were considered. Both of these are measures of a pitchers ability to change speeds. However, at no point was a pitchers willingness to change speeds considered. That would entail looking at how frequently pitches of different speeds are actually thrown.

In several parts of this study it was consistent that the harder a pitcher can throw the lower his expected ERA becomes. It is also consistent that knowing how hard a pitcher can throw his fastball is enough information to predict ERA in terms of velocity, as the speed of all other pitches are highly correlated with fastball velocity.

One inconsistency in this study is the relationship between the range of fastball velocity and ERA. Looking at pitchers in 2020, it appeared the greater the range the lower the pitcher projected ERA would be. However, this was largely influenced by a handful six of outlier observations with larger than average ranges. When comparing to pitchers in 2019, this relationship flipped, and was again due to a handful of this time 5 outlier observations. Due to the inconsistency of this relationship between the two years, I would say there is inconclusive statistical evidence in this study that the range of fastball velocity has any relationship with a pitchers ERA.

Overall, there is much more to pitching than how hard a pitcher throws. How well can he control his pitches? How much do his pitches move? How repeatable and/or deceptive is his delivery? However, when looking exclusively at velocity, how fast a pitcher throws matters. How much he can change speeds doesn't seem to matter as much if at all in this study.

So how will you respond to these clichés? Here are my answers:

"These guys today are just throwers, not pitchers."

That may not be a bad thing. If it comes at the cost of pitch accuracy or other issues that is a separate conversation, but strictly in terms of velocity I'll take the guy that can throw harder without caring so much about the velocity separation of his pitches.

"You have to change speeds to throw hitters off."

This is inconclusive. It might not matter so much if a pitchers changeup is 6 mph slower than his fastball or 15 mph slower, but how frequently he uses it could matter.

"The slower your changeup is, the faster your fastball is."

There is no statistical evidence for this. The faster your fastball is, the faster your fastball is.